Journal of Global Research in Electronics and Communication (JGREC)
Volume 2 | Issue 10 | October 2026 | Pages 23–28
DOI: https://doi.org/xxxxxxxxxxxxx
e-ISSN: 2321-3175
REVIEW ARTICLE OPEN ACCESS Available online: https://jgrec.info/index.php/jgrec/index

Edge-Cloud Continuum for Generative AI: Architectures, Challenges, and Future Directions

Vivek Sharma
System Analyst
MITS Deemed University
VIVEKS@MITSGWALIOR.IN

Corresponding Author: * Vivek Sharma (VIVEKS@MITSGWALIOR.IN)

Received: 15 October 2026 | **Revised:** 20 October 2026 | **Accepted:** 28 October 2026

***Abstract*—A rapidly developing Generative AI (GenAI) is pushing for computing infrastructures capable of handling large-scale inference and training outside of the centralized cloud system limits. The Edge-Cloud Continuum is a very attractive structure for distributing generative workloads throughout a series of heterogeneous compute tiers - starting from very limited resource edge devices to very powerful cloud clusters. This review paper provides a comprehensive survey of architectures, enabling technologies, deployment challenges, and application domains relevant to Generative AI integration within the Edge-Cloud Continuum. The paper explores different types of AI architecture. Also, it considers three ways to optimize models using compression, quantization, or pruning, knowledge distillation, as well as federated learning and hardware acceleration; all these are considered as optimization techniques. Besides that, the study discusses in detail the main challenges and comprehensively examines the real-world applications of GenAI within the Edge-Cloud Continuum, and also proposes future research directions that could advance the development of seamless, intelligent, and efficient edge-cloud generative systems.**

***Keywords*—Gen AI, Edge-Cloud Continuum, LLMs, Federated Learning, Model Compression, Edge Computing, Distributed Inference, Knowledge Distillation, Resource Optimization, Latency.**

I. Introduction

Generative AI (GenAI) is one of the most advanced areas of AI development, allowing machines to produce human-like text or images, audio, video and even code [1][2]. With the rapid growth of foundation models like Large Language Models (LLMs) or multimodal systems [3], the applications for AI solutions have greatly expanded into different areas, such as healthcare, education, scientific research, creation of content, etc [4][5].

Many of the capabilities provided by generative AI are currently available in cloud deployments due to the large number of computational and storage resources required to run such solutions [6]. However, the challenge with deploying these solutions on the cloud is that many of them have performance issues due to latency, communication overhead, and a lack of privacy protection, as well as the reliance on a stable network connection for proper functioning [7]. All of these limitations create barriers to the successful operation and responsiveness of real-time, mission-critical applications that require on-the-ground processing and the ability to respond quickly to user inputs.

These challenges of the Edge Cloud Continuum is being created as a computing paradigm that integrates edge devices, intermediate compute resources, and cloud infrastructure into a unified compute environment [8]. By distributing intelligence across layers of the computing architecture, incorporating low-latency processing, more efficient resource utilization, and improved scalability for Generative AI Applications, the Edge-Cloud Continuum has the opportunity to become an integral part of the future of IT.

A. Structure of the Paper

The paper is organized as follows: Section II presents Generative AI fundamentals, Section III discusses edge–cloud architectures, Section IV highlights key challenges, Section V covers applications and performance metrics, Section VI reviews related work, and Section VII concludes with future directions.

II. Fundamentals of Generative AI

Generative Artificial Intelligence (Generative AI) refers to AI systems capable of generating new content, including text, images, audio, video, and code, by learning patterns from large datasets. Recent advances in deep learning, transformer architectures, diffusion models, and multimodal learning have significantly expanded the capabilities of generative systems. This section presents the key concepts underlying modern Generative AI technologies.

A. Evolution of Generative AI

The evolution of Generative AI has progressed from traditional statistical and probabilistic models to advanced deep learning-based architectures capable of generating realistic and diverse content. Continuous developments in neural networks, transformers, and diffusion models have significantly enhanced the performance and applicability of generative systems across various domains.

1) Early Generative Models

Copyright © 2026 The Author(s).
This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, adaptation, and reproduction in any medium, provided the original author(s) and the source are properly credited.

J. Glob. Res. Elec. Comm., 2 (10):23-28
DOI: https://doi.org/xxxxxxxxxxxx

The development of generative AI began with statistical and probabilistic approaches that relied on handcrafted probability distributions. Later, neural network-based methods such as Restricted Boltzmann Machines (RBMs) and Deep Belief Networks (DBNs) enabled more effective data representation and generation.

2) Deep Learning-Based Generation

The introduction of Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) significantly improved generative capabilities. These models enabled realistic image and video synthesis while providing robust probabilistic representations of complex data.

3) Transformer and Multimodal Era

Recent advances in transformer-based architectures, diffusion models, and multimodal systems such as GPT, CLIP, and DALL-E have further enhanced content generation across diverse applications [9].

B. Large Language Models (LLMs)

LLMs (Large Language Models) utilize deep learning to create an understanding of human language based on the large amount of text that that are trained on. Today's most advanced LLMs are built using the Transformer type of model, which uses a self-attention method to help models capture longer dependencies in a given piece of text [10]. This type of model can accomplish a wide variety of things, such as language translation, creating summated text, answering questions, and conversational AI, etc. Models of this type have improved greatly in their ability to perform tasks by being scaled up in their number of parameters and their amount of training data, that are very effective in many applications.

1) Transformer Architecture

LLMs are built primarily on Transformer architectures that utilize self-attention mechanisms to capture long-range dependencies in text.

2) Capabilities of LLMs

These models support tasks such as language translation, text summarization, question answering, and conversational AI. Scaling model parameters and training data has significantly improved their performance across various applications.

C. Diffusion Models

Diffusion models have emerged as the leading paradigm for high-quality generative image and video synthesis. They operate by gradually adding Gaussian noise to training data and learning a reverse denoising process parameterized by a deep neural network, typically a U-Net or transformer backbone. At inference time, the model iteratively denoises a random noise sample to produce a coherent output [11] Score-based diffusion models and denoising diffusion probabilistic models (DDPMs) have demonstrated remarkable fidelity in image generation and have been extended to audio synthesis, molecule design, and video generation. Latent diffusion models (LDMs), which operate in a compressed latent space rather than pixel space, significantly reduce the computational load and have enabled practical deployment in production systems like Stable Diffusion [12].

1) Working Principle

Diffusion models generate content by gradually adding noise to training data and learning a reverse denoising process. During inference, they iteratively transform random noise into meaningful outputs.

2) Latent Diffusion Models

Latent Diffusion Models (LDMs) operate in compressed latent spaces, reducing computational requirements and enabling practical deployment in systems such as Stable Diffusion.

D. Multimodal Generative AI

Multimodal Generative Intelligent Agents build on language and diffusion models by enabling them to perform tasks across multiple data modalities simultaneously [13]. Multimodal models learn joint representations that enable them to create, translate and reason about multiple modalities within the same framework. For example, Computer systems such as GPT-4V, Gemini Ultra and LLaVA can accept inputs in both text and images and produce coherent outputs that demonstrate a strong understanding of how the two modalities relate.

1) Multimodal Learning

Multimodal models learn joint representations across different data modalities, enabling unified reasoning and content generation.

2) Representative Models

Systems such as GPT-4V, Gemini Ultra, and LLaVA can process both text and image inputs to produce coherent multimodal outputs.

3) Edge-Cloud Challenges

In Edge-Cloud environments, multimodal systems require real-time processing of data from cameras, microphones, and sensors, increasing demands on computation, memory, and communication bandwidth.

III. Architectures for Generative AI in the Edge-Cloud Continuum

The architectural models governing edge and cloud computing integration define the distribution of computational tasks, data flow, and communication among various entities [14]. These architectures address the trade-offs between latency, computational power, bandwidth consumption, and energy efficiency. These architectural paradigms differ in how computational resources, AI models, and data processing tasks are distributed across edge devices, intermediate nodes, and cloud infrastructure.

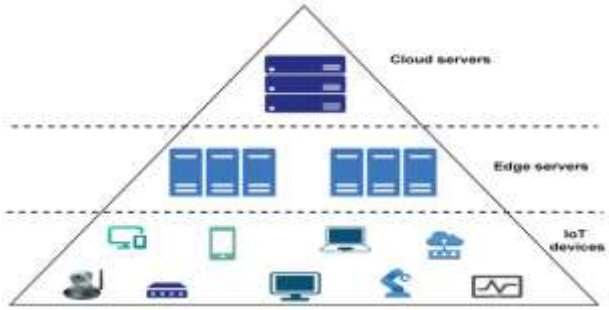

Fig. 1. Schematic Representation of Edge Cloud Architecture

The following Fig 1 illustrates how the compute tiers relate across the full continuum and how a well-structured architectural framework is important to achieving optimal service delivery in applications.

DOI: https://doi.org/xxxxxxxxxxxx

A. Cloud-Centric

Cloud-centric architectures perform all generative AI computation in large-scale cloud-based data centers equipped with GPU or TPU clusters. Edge devices function as thin clients that send user inputs, such as text, speech, or images, to the cloud, receive generated responses, and render the results. Cloud computing provides virtually unlimited horizontal scalability along with centralized model governance, versioning, and monitoring [15]. This approach is commonly used for deploying frontier models such as GPT-4o and Gemini Ultra.

B. Edge-Centric

Edge-centric architectures move generative inference as close as possible to the data source, including smartphones, embedded processors, and autonomous system controllers [16]. The primary objectives are to eliminate network dependency for latency-critical tasks and to keep sensitive user data entirely on-device. This approach has gained significant momentum as modern mobile SoCs integrate dedicated Neural Processing Units (NPUs) capable of executing quantized LLMs containing billions of parameters [17].

C. Hybrid Edge-Cloud

Hybrid Edge-Cloud architectures dynamically route generative workloads between edge and cloud environments based on task complexity, network conditions, and privacy requirements [18].

Key characteristics include:

- Lightweight tasks such as intent detection and preprocessing are executed locally.
- Complex reasoning and large-context generation are offloaded to cloud servers.
- Routing decisions are updated dynamically according to resource availability, battery status, and user preferences.

D. Hierarchical Multi-Tier

Hierarchical Multi-Tier architectures organize computing resources across multiple levels, including end devices, gateways, edge servers, fog nodes, and cloud infrastructure. Tasks are assigned to the tier that best matches their computational and latency requirements.

Typical responsibilities of each tier include:

- End devices perform data collection and basic processing.
- Edge and fog nodes handle preprocessing and latency-sensitive inference.
- Cloud servers execute large-scale model training, orchestration, and long-term storage.
- This layered architecture improves scalability, resource utilization, and service efficiency across the Edge-Cloud Continuum.

E. Distributed Inference

Distributed Inference architectures decompose a single large model across multiple networked nodes, enabling models that exceed the capacity of any individual device to execute collaboratively [19]. Pipeline parallelism assigns successive transformer layers to different nodes, while tensor parallelism partitions individual layers for concurrent execution. Activation transfer between nodes remains the primary bottleneck in bandwidth-constrained environments.

F. Collaborative AI

Collaborative AI architectures provide coordinated operation among multiple AI models deployed across different tiers of the continuum. Lightweight on-device models perform immediate processing, context gathering, and tool invocation, while larger cloud-hosted models handle complex reasoning and knowledge-intensive tasks [20]. Dynamic routing mechanisms determine which model processes a request based on latency requirements, resource availability, and task complexity. This architecture is particularly beneficial for agentic AI applications.

IV. Challenges in Edge-Cloud Generative AI

The deployment of Generative AI across the Edge-Cloud Continuum introduces several technical challenges related to resource limitations, communication efficiency, latency requirements, energy consumption, and data security. As generative models continue to increase in size and complexity, addressing these challenges becomes essential for achieving efficient, scalable, and reliable AI services in distributed environments.

A. Computational Resource Constraints

Edge devices typically have insufficient memory and processing power for most large generative models, which is evidenced by the resource requirements of a quantized model.

B. Latency and Real-Time Processing

Due to the latency introduced by network round-trip delays that occur when using cloud-based inference services, which is not an appropriate option for many real-time applications that require a response time of less than 200 milliseconds.

C. Communication Overhead

The transfer of raw input data, activation data for intermediates, and model updates at multiple tiers in the processing system causes a significant amount of bandwidth to be consumed and further adds latency to the overall pipeline of the processing system.

D. Energy Consumption

The time required to perform large language model inference continuously will quickly consume battery resources in smartphones, thus making energy efficiency a key limitation in edge device deployment.

E. Privacy and Security

Transferring sensitive personal data to cloud-based services creates potential privacy violations under regulations such as HIPAA and the GDPR, which is a primary motivator for using edge computing technologies [21].

F. Model Management and Updating

Generative AI models require frequent updates to improve performance, incorporate new knowledge, and address security vulnerabilities. Managing model synchronization across distributed edge and cloud environments can be challenging due to heterogeneous hardware capabilities, storage limitations, and varying network conditions.

J. Glob. Res. Elec. Comm., 2 (10):23-28
DOI: https://doi.org/xxxxxxxxxxxx

G. *Scalability and Heterogeneity*

Edge-Cloud environments consist of diverse devices with different computational capacities, operating systems, and network characteristics. Ensuring consistent performance and scalable deployment of generative AI services across such heterogeneous infrastructure remains a significant challenge.

V. Applications of Edge-Cloud Generative AI and Performance Evaluation

By distributing intelligence across edge devices and cloud resources, Generative AI can support a wide range of applications.

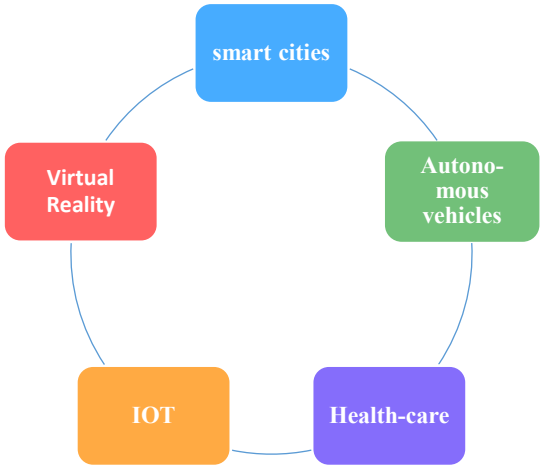

Fig. 2. Applications of EDGE-CLOUD Continuum

Fig 2 illustrates the major application domains of Generative AI within the Edge-Cloud Continuum.

- **Smart Cities:** Edge nodes handle real-time inference for traffic and pedestrian management, while cloud systems generate longer-horizon urban planning reports. Generative AI synthesizes actionable operator summaries from heterogeneous sensor streams without exposing raw surveillance data.
- **Autonomous Vehicles:** Safety-critical perception and planning loops run entirely on-board, while cloud connectivity supports model updates and fleet-level federated learning [22]. Generative models contribute to trajectory planning, synthetic training data generation, and passenger interaction interfaces.
- **Healthcare Systems:** Cloud-side LLMs support clinical documentation, diagnosis assistance, and multimodal radiology report generation.
- **Industrial IOT:** Generative AI produces natural language maintenance reports and root cause analyses, which are offloaded to cloud servers during off-peak periods, by performing real time detection.
- **Augmented and Virtual Reality:** Cloud services provide high-fidelity asset generation and persistent multi-user state during natural pauses in user interaction [23].

A. *Performance Evaluation Metrics*

Performance evaluation metrics are essential for assessing the efficiency, scalability, and effectiveness of Generative AI systems deployed across Edge-Cloud environments. These metrics measure system responsiveness, resource usage, energy consumption, throughput, and output quality.

- **Latency:** Latency is measured as time to first token (TTFT) for responsiveness and inter-token latency (ITL) for output stream smoothness. Evaluations should report p50, p95, and p99 percentiles to capture real-world tail behavior across varying network conditions.
- **Throughput:** Throughput numbers, given in tokens per second, give an idea of how many users a system can handle simultaneously with a certain hardware budget [24].
- **Energy Efficiency:** Energy efficiency, which is measured in tokens per joule, is very important In particular when the inference is performed on battery-powered edge devices [25].
- **Resource Utilization:** Monitoring GPU, memory, and bandwidth utilization offers an insight into scheduling bottlenecks and capacity wastage across tiers.
- **Quality of Generated Output:** It measures for evaluating text quality are perplexity BLEU ROUGE, and BERT Score, whereas for image quality, it is FID, CLIP score, and LPIPS [26].

VI. Literature Review

The literature review highlights advance in edge–cloud continuum architectures, distributed intelligence, and generative AI integration, emphasizing optimization, orchestration, and digital twin applications. It also identifies key challenges including latency, interoperability, scalability, and infrastructure heterogeneity across IoT-driven distributed computing environments.

L. BELCASTRO et al.(2026) Driven by the exponential growth of IoT-generated data and the need for real-time responsiveness, the edge-cloud continuum features multi-layered architectures. However, its adoption is hindered by infrastructural challenges, fragmented standards, and limited guidance. This survey addresses these gaps by introducing a developer-oriented conceptual framework and examining architectural models, performance metrics, computation, communication, deployment paradigms, enabling technologies, edge-to-cloud platforms, real-world applications, and testing tools [27].

J. Liu et al. (2026) This survey comprehensively examines distributed intelligence and model optimization in edge-cloud environments, covering architectures, communication protocols, network-aware computing frameworks, model optimization approaches, AI-driven resource management, privacy and security aspects, practical applications, and performance analysis and benchmarking techniques for evaluating complex distributed systems [28].

P. K. Myakala (2025) This survey presents a comprehensive review of Distributed and Decentralized Intelligence (DDI) systems, covering their architectural foundations, computational models, orchestration mechanisms, virtualization technologies, federated learning, lightweight models, hierarchical intelligence, collaborative AI approaches, and applications in smart healthcare, autonomous systems, industrial IoT, and smart cities [29].

L. Belcastro, F. Marozzo, and A. Presta (2025) This paper presents a generative AI framework that leverages Large Language Models (LLMs) to optimize service mapping across the edge–cloud continuum by automatically matching abstract

DOI: https://doi.org/xxxxxxxxxxxx

application components to suitable edge and cloud services while satisfying functional, nonfunctional, and QoS requirements across heterogeneous infrastructures [30].

C. Savaglio et al.(2024) The rise of IoT and AI has enabled Digital Twins (DT) to exchange real-world data and utilize it for achieving their objectives. Recently, Generative AI (Gen-AI) has been applied to DT in various contexts. This article defines Generative DT (GDT), highlighting the role of Gen-AI in integrating model-driven and data-driven approaches for DT development. It further analyzes the value of Gen-AI architectures in enhancing DT performance within IoT environments and edge-cloud continuum deployments [31].

Y.-C. Wang, J. Xue, C. Wei, and C.-C. J. Kuo (2023) The rapid development of GenAI systems has generated large volumes of Internet data, creating challenges for existing computing and communication frameworks. While current GenAI services depend on cloud computing for high computational resources, they often suffer from high latency due to data transmission and heavy user demand. Edge-cloud computing offers a promising alternative by combining edge and cloud resources to achieve both low latency and sufficient computation power. This overview reviews recent advances in GenAI and edge-cloud computing and discusses key challenges in scaling GenAI applications using edge-cloud collaborative systems [32].

The Table I compare edge-cloud generative AI studies, highlighting architectures, distributed intelligence, LLM optimization, digital twins, challenges and future directions of research.

TABLE I. COMPARATIVE ANALYSIS OF RECENT STUDIES ON EDGE-CLOUD GENERATIVE AI

Authors	Study On	Approach	Key Findings	Challenges / Limitations	Future Directions
Belcastro et al. (2026)	Edge–cloud continuum frameworks	Conceptual developer-oriented survey	Unified view of architecture, metrics, tools, and platforms	Lack of standardization, fragmented tools, deployment complexity	Standard frameworks and unified orchestration models
Liu et al. (2026)	Distributed intelligence in edge-cloud	Survey of AI + network-aware optimization	Covers optimization, security, protocols, and benchmarking	High system complexity, limited real-world validation	Scalable AI-driven resource management and security-aware systems
Myakala (2025)	Distributed & decentralized intelligence (DDI)	Comprehensive review	Highlights federated learning and hierarchical AI models	Communication overhead and limited interoperability	Lightweight decentralized models and improved orchestration
Belcastro et al. (2025)	LLM-based service mapping	Generative AI framework	LLMs can optimize service selection across edge-cloud layers	QoS constraints handling and scalability issues	Adaptive LLM-based orchestration for heterogeneous systems
Savaglio et al. (2024)	Generative AI in Digital Twins	Gen-AI enabled DT architecture	GenAI improves DT performance in IoT-edge-cloud systems	Integration complexity and real-time synchronization issues	Scalable Generative Digital Twin (GDT) systems
Wang et al. (2023)	GenAI + edge-cloud computing	System overview	Edge-cloud reduces latency and improves scalability	Network bottlenecks and resource imbalance	Efficient GenAI distribution and latency-aware architectures

VII. CONCLUSION AND FUTURE WORK

Generative AI has emerged as a transformative technology, and its integration with the Edge–Cloud Continuum presents a powerful paradigm for distributed intelligent computing systems. This paper systematically reviewed foundational concepts, architectural models, applications, challenges, and recent advancements in this domain. The study demonstrates that combining edge and cloud resources significantly enhances system performance by reducing latency, improving scalability, optimizing resource utilization, and strengthening data privacy for generative AI workloads. However, several challenges remain, including limited computational capacity at the edge, high communication overhead, system heterogeneity, and the need for efficient real-time orchestration. The comparative literature analysis further indicates that no single architecture is sufficient to address all requirements, highlighting the importance of hybrid and adaptive frameworks. Overall, the Edge–Cloud Continuum provides a strong foundation for next-generation Generative AI systems across diverse application domains.

Future research should focus on lightweight model optimization, intelligent task offloading, standardized orchestration frameworks, energy-efficient inference mechanisms, and secure multi-tier deployment strategies to enable scalable real-world adoption.

- **Funding:** This research received no external funding.
- **Institutional Review Board Statement:** Not applicable.
- **Informed Consent Statement:** Not applicable.
- **Data Availability Statement:** The data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Conflicts of Interest:** The author declares no conflict of interest.

How to cite this article:
V. Sharma, "Edge-Cloud Continuum for Generative AI: Architectures, Challenges, and Future Directions," *Journal of Global Research in Electronics and Communication*, vol. 2, no. 10, pp. 23–28, Oct. 2026. Doi: XXXX

REFERENCES

[1] M. Parikh, A. A. Soni, S. M. Shah, and A. R. Jha, "Big Data Workload Profiling for Energy-Aware Cloud Resource Management," Jan. 2026.

[2] S. Yazdani *et al.*, "Generative AI in Depth: A Survey of Recent Advances, Model Variants, and Real-World Applications," *J. Big Data*, vol. 12, no. 1, p. 230, Oct. 2025, doi: 10.1186/s40537-025-01247-x.

[3] T. P. Patel, "Adaptive Token Routing for Heterogeneous LLM Inference in Edge-Cloud Continuum," in *SoutheastCon 2026*, Huntsville, AL, USA: IEEE, 2026, pp. 1–7, April. doi: 10.1109/SoutheastCon63549.2026.11476596.

[4] P. Kumar, "Edge Computing and IoT for Real-Time Healthcare Data Processing and Integration," in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Dec. 2025, pp. 105–110. doi: 10.1109/ICAAIC64647.2025.11331211.

[5] P. Kumar, "Leveraging Generative AI for Automated Data Standardization and Interoperability in Healthcare," in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Dec. 2025, pp. 99–104. doi: 10.1109/ICAAIC64647.2025.11330217.

[6] N. Bhandarwar, "Integrating Generative AI and Model Context Protocol (MCP) with Applied Machine Learning for Advanced Agentic AI Systems," *Int. J. Comput. Trends Technol.*, vol. 73, no. 8, pp. 7–14, August, 2025, doi: 10.14445/22312803/IJCTT-V73I8P102.

[7] M. Navardi *et al.*, "GenAI at the Edge: Comprehensive Survey on Empowering Edge Devices," *Qeios*, vol. 7, no. 6, pp. 1–8, Feb, Jun. 2025, doi: 10.32388/JEU3U0.2.

[8] D. Rosendo, A. Costan, P. Valduriez, and G. Antoniu, "Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review," *J. Parallel Distrib. Comput.*, vol. 166, pp. 71–94, Apr. 2022, doi: 10.1016/j.jpdc.2022.04.004.

[9] M. Trigka and E. Dritsas, "The Evolution of Generative AI: Trends and Applications," *IEEE Access*, vol. 13, no. May, pp. 98504–98529, 2025, doi: 10.1109/ACCESS.2025.3574660.

[10] X. Ren, J. Tang, D. Yin, N. Chawla, and C. Huang, "A Survey of Large Language Models for Graphs," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2024, pp. 6616–6626. doi: 10.1145/3637528.3671460.

[11] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Adv. Neural Inf. Process. Syst.*, vol. 2020-Decem, pp. 6840–6851, Dec. 2020, doi: 10.48550/arXiv.2006.11239.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, vol. 2022-June, pp. 10674–10685, Apr. 2022, doi: 10.1109/CVPR52688.2022.01042.

[13] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. Mach. Learn. Res.*, vol. 139, pp. 8748–8763, Feb. 2021, doi: 10.48550/arXiv.2103.00020.

[14] C. Chen and C. Liu, "A 3 . 5-tier container-based edge computing architecture," *Comput. Electr. Eng.*, vol. 93, no. August 2020, p. 107227, 2021, doi: 10.1016/j.compeleceng.2021.107227.

[15] Q. Wu *et al.*, "AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation," pp. 1–46, Oct. 2023, doi: 10.48550/arXiv.2308.08155.

[16] J. Lin *et al.*, "Awq : A Ctivation - Aware W Eight Q Uantization For O N -D Evice Llm C Ompression and A Cceleration," 2022.

[17] B. Krishnan, S. Perla, S. Maddela, and R. Lingam, "Adaptive Multi-Cloud Infrastructure for CRM Analytics: Real-Time ML and Data Sync with LLMs," in *2025 IEEE 3rd Global Conference on Wireless Computing and Networking (GCWCN)*, Lonawala, Maharashtra, India: IEEE, Nov. 2025, pp. 1–8. doi: 10.1109/GCWCN66157.2025.11448404.

[18] Y. He, J. Fang, F. R. Yu, and V. C. Leung, "Large Language Models (LLMs) Inference Offloading and Resource Allocation in Cloud-Edge Computing: An Active Inference Approach," *IEEE Trans. Mob. Comput.*, vol. 23, no. 12, pp. 11253–11264, 2024, doi: 10.1109/TMC.2024.3415661.

[19] R. Y. Aminabadi *et al.*, "DeepSpeed- Inference: Enabling Efficient Inference of Transformer Models at Unprecedented Scale," in *SC22: International Conference for High Performance Computing, Networking, Storage and Analysis*, IEEE, Nov. 2022, pp. 1–15. doi: 10.1109/SC41404.2022.00051.

[20] Z. Xi *et al.*, "The rise and potential of large language model based agents: a survey," *Sci. China Inf. Sci.*, vol. 68, no. 2, p. 121101, 2025, doi: 10.1007/s11432-024-4222-0.

[21] F. Mireshghallah, K. Goyal, A. Uniyal, T. Berg-Kirkpatrick, and R. Shokri, "Quantifying Privacy Risks of Masked Language Models Using Membership Inference Attacks," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 8332–8347. doi: 10.18653/v1/2022.emnlp-main.570.

[22] J. Yang *et al.*, "Generalized Predictive Model for Autonomous Driving," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2024, pp. 14662–14672. doi: 10.1109/CVPR52733.2024.01389.

[23] Y. Liang *et al.*, "Latency Reduction in Immersive Systems through Request Scheduling with Digital Twin Networks in Collaborative Edge Computing," 2024.

[24] C. Holmes *et al.*, "DeepSpeed-FastGen: High-throughput Text Generation for LLMs via MII and DeepSpeed-Inference," vol. 1, pp. 1–14, Jan. 2024, doi: 10.48550/arXiv.2401.08671.

[25] N. Bannour, S. Ghannay, A. Névéol, and A.-L. Ligozat, "Evaluating the carbon footprint of NLP methods: a survey and analysis of existing tools," in *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 11–21. doi: 10.18653/v1/2021.sustainlp-1.2.

[26] L. Zheng *et al.*, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," in *Advances in Neural Information Processing Systems 36*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., San Diego, California, USA: Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023, pp. 46595–46623. doi: 10.52202/075280-2020.

[27] L. Belcastro, F. Marozzo, A. Orsino, D. Talia, and P. Trunfio, "Navigating the Edge-Cloud Continuum: A State-of-Practice Survey," *IEEE Access*, vol. 14, pp. 40622–40647, May 2025, doi: 10.1109/ACCESS.2026.3673012.

[28] J. Liu *et al.*, "Edge-Cloud Collaborative Computing on Distributed Intelligence and Model Optimization: A Survey," *IEEE Commun. Surv. Tutorials*, vol. 28, pp. 5049–5080, 2026, doi: 10.1109/COMST.2026.3669216.

[29] P. K. Myakala, "Distributed Data Intelligence in the Edge-to-Cloud Continuum: a Survey of Architectures, Models, and Applications," in *2025 6th International Conference on Data Intelligence and Cognitive Informatics (ICDICI)*, Tirunelveli, India: IEEE, Jul. 2025, pp. 93–99. doi: 10.1109/ICDICI66477.2025.11135068.

[30] L. Belcastro, F. Marozzo, and A. Presta, "Generative AI for Optimizing Service Mapping in the Edge-Cloud Continuum," in *2025 IEEE Conference on Pervasive and Intelligent Computing (PICom)*, Hakodate, Japan: IEEE, 2025, pp. 52–57, October. doi: 10.1109/PICom68402.2025.00013.

[31] C. Savaglio, V. Barbuto, F. Mangione, and G. Fortino, "Generative Digital Twins: A Novel Approach in the IoT Edge-Cloud Continuum," *IEEE Internet Things Mag.*, vol. 8, no. 1, pp. 42–48, August, 2024, doi: 10.1109/IOTM.001.2400035.

[32] Y.-C. Wang, J. Xue, C. Wei, and C.-C. J. Kuo, "An Overview on Generative AI at Scale With Edge–Cloud Computing," in *IEEE Open Journal of the Communications Society*, IEEE, 2023, pp. 2952–2971, October. doi: 10.1109/OJCOMS.2023.3320646.